Welk percentage gezakte studenten is bij een bepaalde toetsvorm nog acceptabel? De mening van NVMO-congresgangers

J. Cohen-Schotanus, J. Schönrock-Adema, A.J.J.A. Scherpbier

Samenvatting

Inleiding: *De slaagpercentages voor toetsen variëren nogal eens, zowel binnen eenzelfde vak als tussen vakken. Deze variatie heeft onder andere te maken met het feit dat de toets de ene keer moeilijker is dan de andere. Een mogelijke manier om deze variaties te beperken is om de cesuur aan te passen. Er bestaat geen 'perfecte' methode om de cesuur te bepalen. Wel moet een cesuurmethode aan een aantal voorwaarden voldoen. Een van die voorwaarden is dat de uitslag acceptabel moet zijn. Onderdeel van een acceptabele uitslag is het percentage gezakte studenten. Wij hebben onderzocht welk percentage gezakte studenten bij een bepaalde toetsvorm nog acceptabel is voor de medische opleiding.*

Methode: *De deelnemers aan het NVMO-congres 2004 zijn gevraagd om voor negen toetsvormen aan te geven wat een nog acceptabel percentage gezakte studenten is. Er is een subgroep onderscheiden van 'medical education experts'.*

Resultaten: *De deelnemers aan het NVMO-congres vinden dat bij de schriftelijke toets in de pre-klinische fase de meeste studenten mogen zakken (26%), bij de gedragsbeoordeling in de klinische fase de minste (10%). De 'medical education experts' lijken iets milder in hun oordeel te zijn dan de rest van de congresgangers.*

Conclusie: *De congresgangers doen een duidelijke uitspraak over het nog acceptabele percentage gezakte studenten bij de verschillende toetsvormen. Faculteiten zouden er goed aan doen zelf grenzen te stellen en aan te geven wat te doen bij overschrijding. (Cohen-Schotanus J, Schönrock-Adema J, Scherpbier A.J.J.A. Welk percentage gezakte studenten is bij een bepaalde toetsvorm nog acceptabel? De mening van NVMO-congresgangers. Tijdschrift voor Medisch Onderwijs 2005;24(4):184-189.)*

Inleiding

Een onverwacht hoog percentage gezakte studenten bij een toets is zowel voor de docent als de student vrijwel altijd vervelend. Veel docenten hebben bij een desastreuze uitslag de neiging iets met de cesuur van hun toets te doen. Hoe de cesuur van toetsen bepaald moet worden, is een onderwerp dat steeds op de onderwijsagenda staat. In de literatuur zijn verschillende methodes voor cesuurbepaling te vinden.[1-6] Nederlandse docenten hebben een sterke voorkeur voor een absolute cesuur: de student die, indien nodig na correctie voor raden, 60% van de toetsvragen goed beantwoordt, verdient een '6'. Deze historische manier van cesuurbepaling blijkt niet te voldoen, omdat de uitslagen te vaak inconsistent zijn met de verwachtingen.[7] Er is en wordt dan ook naarstig gezocht naar goede alternatieve methoden. De verschillende cesuurmethoden die de laatste decennia ontwikkeld zijn, kunnen worden verdeeld in drie categorieën. De eerste categorie bestaat uit de cesuur die wordt vastgesteld voordat de toets wordt afgenomen, ook wel de absolute methode genoemd. Voorbeelden van deze manier van cesuurbepaling zijn de methoden van Ebel en Angoff.[1,3] Bij het vaststellen van

deze cesuren wordt gewerkt met expertpanels. Panelmethoden zijn echter duur en het lukt meestal niet deze procedures in de dagelijkse praktijk te realiseren. De tweede categorie van cesuurmethoden bestaat uit de cesuur die na toetsafname wordt vastgesteld en gebaseerd is op de prestatie van de groep (relatieve methode). Een goed voorbeeld van een dergelijke cesuur is de cesuur van Wijnen.[5] Tot slot zijn er methoden waarbij de absolute en relatieve methoden gecombineerd worden, zoals bijvoorbeeld de methoden van Hofstee en Cohen-Schotanus et al.[4,6]

Wanneer bij eenzelfde toets verschillende cesuren toegepast worden, blijkt dat het percentage geslaagde studenten bij de ene methode (veel) hoger is dan bij de andere methode.[8-10] Welke cesuur bij die toets de 'juiste' is, is afhankelijk van de verwachtingen die de betrokken docenten hebben betreffende het beheersingsniveau van de studenten, het slaagpercentage en het belang van de toets in een examenprogramma. Een gouden standaard voor het vaststellen van de cesuur is dan ook niet te geven. Er bestaan geen 'perfecte' of 'foute' cesuren of volledig objectieve methoden waar alle subjectieve invloeden uit verwijderd zijn. Wel zijn er in de loop der jaren een aantal kenmerken geformuleerd waaraan een cesuur moet voldoen:[7]

- De betrokkenen moeten de (kwaliteit van de) toets beoordelen en het eens zijn over de te hanteren methode.
- De methode moet geloofwaardig zijn.
- De methode moet gebaseerd zijn op onderzoek.
- De methode moet eenvoudig te begrijpen en uit te voeren zijn (transparantie).
- De methode moet een realistische uitkomst opleveren.

Met een realistische uitkomst wordt bedoeld een geloofwaardig niveau en een acceptabel percentage geslaagden. Het is dit laatste aspect waar dit artikel op ingaat. Wanneer is er eigenlijk sprake van een acceptabel percentage geslaagden?

In de literatuur is weinig te vinden over 'kritieke' percentages gezakte studenten. Alleen Stern et al. meldden de bevindingen van een internationaal panel van 'medical education experts' betreffende het percentage gezakte studenten dat men tolerabel vond. Dit percentage varieerde afhankelijk van de toetsvorm van 9% (observaties) tot 24% (multiple choice toetsen).[11] Wij vroegen ons af wat de deelnemers aan het jaarlijkse congres van de Nederlandse Vereniging van Medisch Onderwijs (NVMO) een acceptabel percentage gezakte studenten vinden.

De vragen die in dit onderzoek aan de orde gesteld worden, zijn:

- Wat is voor personen die betrokken zijn bij het onderwijs aan studenten geneeskunde een net nog acceptabel percentage gezakte studenten en is dit percentage afhankelijk van de toetsvorm?
- Hebben 'medical education experts' hierover andere opvattingen dan 'niet-medical education experts'?

Methode

Onderzoeksgroep

Het onderzoek is uitgevoerd onder de deelnemers aan het jaarlijkse congres van de NVMO in november 2004. Voor dit congres hadden zich ruim 500 personen ingeschreven. De vragenlijst is tijdens de eerste plenaire sessie van het congres ingevuld. Deze sessie werd door ruim 400 personen bijgewoond. Een van de auteurs heeft de bedoeling van het onderzoek toegelicht.

Instrument

Aan de respondenten is gevraagd voor negen verschillende toetsvormen aan te geven wat zij een nog acceptabel percentage

gezakte studenten zouden vinden. Daarbij werd expliciet gevraagd naar de eigen mening en niet naar een eventuele opvatting in de werkomgeving. Tevens is gevraagd om een aantal persoonskenmerken in te vullen betreffende functie, gender, lidmaatschap NVMO, plaats van herkomst en in hoeverre men zichzelf beschouwt als 'opinion leader'. De onderzoekers hebben 'medical education expert' op basis van de persoonskenmerken als volgt gedefinieerd: de personen die werkzaam zijn bij een medische faculteit, meer dan drie jaar lid zijn van de NVMO en zichzelf als 'opinion leader' beschouwen.

Data-analyse

Verschillen tussen 'medical education experts' en de andere respondenten zijn geanalyseerd met multivariate analyse en de Wilcoxon signed-rank toets. Deze laatste toets geeft aan of er sprake is van een trend betreffende de richting van verschillen.

Resultaten

De vragenlijst werd ingevuld door 404 personen. De statistische analyses hebben betrekking op de 376 respondenten die alle vragen beantwoord hebben. Van de respondenten was 22% student, 11% klinisch docent, 9% docent basisvak, 20% medewerker onderwijsontwikkeling, 16% onderwijscoördinator, 7% behoorde bij de leiding van het onderwijsinstituut en 16% kwalificeerde zich 'anders'. Nog niet iedereen bleek lid te zijn van de NVMO: 42% niet lid, 16% was een jaar lid, 9% twee jaar, 16% 3-5 jaar, 9% 5-10 jaar en 8% meer dan 10 jaar. Een derde (35%) van de aanwezigen beschouwde zichzelf als 'opinion leader' in het medisch onderwijs. Er waren meer vrouwen (57%) dan mannen. Zo'n 80% van de aanwezigen werkt of studeert bij een medische faculteit. Aan de definitie van 'medical education expert' werd door 58 deelnemers voldaan (15%).

Bij het aangeven van het nog acceptabele percentage gezakte studenten maakt 95% van de respondenten onderscheid tussen de verschillende toetsvormen. In tabel 1 wordt per toetsvorm weergegeven wat het gemiddelde percentage gezakte studenten is dat nog acceptabel werd gevonden. Bij de schriftelijke toets in de preklinische fase is het percentage gezakte studenten dat nog acceptabel werd gevonden het hoogst:

Tabel 1. *Gemiddelde nog acceptabele percentages gezakte studenten per toetsvorm genoemd door 'medical education experts' (experts) en de rest van de respondenten (niet-experts).*

Toetsvorm	Gemiddelde Niet-expert		Gemiddelde Expert	
	N=318	SD	N=58	SD
Schriftelijke toets pre-klinische fase	26.12	12.53	24.48	7.79
Schriftelijke toets klinische fase	19.77	12.37	16.38	8.44
Mondelinge toets pre-klinische fase	20.06	12.19	17.31	8.15
Stationsexamen	16.65	12.45	14.83	8.04
Patiënt (klinisch) examen	14.22	12.78	11.43	7.13
Portfolio	12.63	11.56	10.95	7.45
Gedragsbeoordeling pre-klinische fase	14.91	12.58	13.09	8.15
Gedragsbeoordeling klinische fase	11.27	12.65	7.95	7.22
Verslag	13.68	12.50	13.40	9.28

25,7%. Het laagst is het percentage acceptabel gezakte studenten bij de gedragsbeoordelingen in de klinische fase: 10,1%.

De 'medical education experts' laten bij alle toetsvormen een lager acceptabel percentage gezakten zien dan de overige respondenten. Het overall-verschil tussen de experts en de overige respondenten is niet significant in de multivariate analyse ($F(9,366)=1.188$, $p=.301$). De Wilcoxon signed-rank test toont aan dat de trend wel significant is ($z=-2.599$, $p=.009$). Bij de schriftelijke toets in de klinische fase en de gedragsbeoordeling in de klinische fase zijn de verschillen het grootst.

Discussie

De deelnemers aan het NVMO-congres 2004 vinden gemiddeld genomen dat het nog acceptabele percentage gezakte studenten, afhankelijk van de toetsvorm, ongeveer tussen de 10% en 25% moet liggen. Ook ligt het nog acceptabele percentage gezakte studenten aan het begin van de studie (preklinische fase) hoger dan in de laatste (klinische) fase. Experts op het gebied van het medisch onderwijs zijn iets milder in hun oordeel dan de overige respondenten. Zij zullen eerder de cesuur willen aanpassen. Het lijkt ons aannemelijk dat het mildere oordeel van de experts te maken heeft met het feit dat zij vaker geconfronteerd zullen zijn geweest met diverse tekortkomingen van toetsen. Zij zullen bij extreme toetsuitslagen dan ook minder snel geneigd zijn de 'schuld' bij studenten te zoeken. Dat er, in tegenstelling tot de Wilcoxon-toets, bij de multivariate analyse geen significante verschillen zijn gevonden, heeft waarschijnlijk te maken met de relatief kleine groep experts en de grote spreiding in de gegevens van de niet-experts.

Onze bevindingen komen overeen met het oordeel van een internationaal panel van experts op het terrein van medisch onderwijs.[11] De procedure die door het internationale panel is gevolgd, verschilde van onze opzet. De panelleden begonnen met een inleidende sessie en bestudeerden vervolgens het toetsmateriaal van acht medische faculteiten (in China). Door discussies kwamen zij tot een gezamenlijke opvatting over het acceptabele percentage gezakte studenten per toetsvorm. Het is opvallend dat de gemiddelde uitkomst van onze vragenlijst vergelijkbaar is. Het internationale panel oordeelde dat het acceptabele percentage gezakte studenten voor multiple choice-toetsen 24% is. De 'medical education experts' in ons onderzoek geven 24,4% aan als acceptabel percentage gezakte studenten voor schriftelijke toetsen in de preklinische fase van de studie. Voor (gedrags-)observaties zijn de uitkomsten 9% (internationaal) en 8% (nationaal).

De setting waarin ons onderzoek werd uitgevoerd was zo dat de congresgangers werden 'overvallen' om aan het onderzoek mee te doen. Men kan zich dan ook afvragen of een dergelijke setting de onderzoeksresultaten heeft beïnvloed. Veertien respondenten (3%) becommentarieerden het onderzoek. Vijf respondenten hadden problemen met de vraagstelling. Zij vonden de vraagstelling te vaag en om die reden onwetenschappelijk. Negen respondenten gingen in op de verantwoording van hun keuzes. Een aantal wees op het belang van de kwaliteit van de toetsing en het belang van goede begeleiding. Anderen benadrukten dat een acceptabel niveau belangrijker is dan het percentage gezakte studenten. Wij menen op basis van deze reacties te mogen concluderen dat de meeste respondenten de vraagstelling goed begrepen hebben.

De vraag is wat deze uitkomsten te betekenen hebben voor de toetspraktijk. Het

zonder meer laten zakken van grote aantallen studenten voor een toets lijkt voor de meeste deelnemers aan het NVMO-congres niet acceptabel te zijn. In de jaren zestig bediscussieert De Groot in zijn boek *Vijven en Zessen* al de problematiek van het 'onterecht' zakken.[12] Hij gaat daarbij in op het 'gebruik van de cijferschaal' en de daarbij gebruikte beslisregels. De Groot betoogt dat het cijfer dat een student krijgt natuurlijk in de allereerste plaats behoort af te hangen van de *geleverde prestatie* (zie hoofdstuk IV: Cijfers over cijfers). Hij zegt daar niet aan te willen tornen. Vervolgens gaat hij in op alle andere factoren die van (negatieve) invloed zijn op het cijfer: de status van het vak, verschillen tussen scholen, de invloed van de docent, de grootte van de groep, et cetera. Ook vindt hij het opmerkelijk dat in hogere studiejaren het percentage gezakte studenten niet afneemt. Hij betoogt dat een afnemend percentage na de voorgaande selectie het meest logische resultaat zou zijn. Het constante percentage doublures (25%) wordt wel de 'Wet van Posthumus' genoemd. Dit percentage wordt in ons onderzoek alleen gevonden bij de schriftelijke toets in de preklinische fase.

De boeken van Dousma et al. en van Van Berkel worden in het Nederlandse hoger onderwijs vaak geraadpleegd bij toetsproblemen.[13-14] In beide boeken wordt ingegaan op het feit dat diverse factoren van invloed zijn op de kwaliteit van toetsen en er wordt nadrukkelijk geadviseerd hier bij de uitslagbepaling rekening mee te houden. Dousma suggereert het gebruik van een 'al te gek'-clausule en Van Berkel meent dat je studenten beter onterecht kunt laten slagen dan onterecht laten zakken. Geen van beiden geeft echter aanwijzingen wanneer en hoe de clausule in werking gesteld moet worden. Over de hoogte van het nog acceptabele percentage gezakte studenten wordt niet gesproken.

De faculteiten waarbij de auteurs werkzaam zijn hebben in het verleden besloten grenzen aan te geven voor het acceptabele percentage gezakte studenten. Zij baseren zich daarbij op de bevinding dat slechte toetsresultaten veel vaker veroorzaakt worden door variaties in de moeilijkheid en kwaliteit van de toetsing en/of de kwaliteit van het onderwijs dan door een verschil in de kwaliteit van jaargroepen.[6 15] Met andere woorden, de kwaliteit van de studentpopulatie is veel stabieler dan de moeilijkheid van de toets. In Groningen is dit grenspercentage op 30% gesteld, het resultaat van onderhandelingen met de examen- en opleidingscommissie eind jaren tachtig. In Maastricht is de grens 16%, gebaseerd op de gemiddelde zakpercentages van voorgaande jaren. Bij de voortgangstoets die vier keer per jaar door alle geneeskundestudenten in Maastricht, Nijmegen, Groningen en Leiden wordt gemaakt is een compromis bereikt: 20%.

Op basis van ons onderzoek menen wij te mogen concluderen dat zakpercentages hoger dan 25% bij schriftelijke toetsen in de preklinische fase niet zonder meer acceptabel zijn. Faculteiten doen er verstandig aan zelf grenzen te stellen en procedures te ontwikkelen wat te doen bij overschrijding van de grenzen. Gezien de grote spreiding in ons onderzoek wat betreft de mening van de respondenten (zelfs bij deze positieve selectie van deelnemers, te weten de congresgangers), denken wij dat er bij dat proces door onze 'medical education experts' nog heel wat missiewerk te verrichten valt.

Literatuur

1. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, editors. Educational Measurement. 2nd ed. Washington DC: American Council on Education; 1971. p. 508-600.
2. Gruyter DNM de. Compromise models for establishing examination standards. Journal of Educational Measurement 1985;22:263-9.

3. Ebel RL. Essentials of educational measurement. 3rd ed. Englewood Cliffs, NJ: Prentice Hall; 1979.
4. Hofstee WKB. Een gebufferde oplossing voor het bepalen van de grens tussen voldoende en onvoldoende. Universiteit en Hogeschool 1982;28:21-8.
5. Wijnen WHFW. Onder of boven de maat: een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen [dissertation]. Groningen: [s.n.]; 1971.
6. Cohen-Schotanus J, Vleuten CPM van der, Bender W. Een betere cesuur bij tentamens. Onderzoek van Onderwijs 1996;25(3):54-5.
7. Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, Vleuten CPM van der, Newble DI, editors. International handbook of research in medical education. Dordrecht [etc.]: Kluwer Academic Publishers; 2002. p. 811-34.
8. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. Acad Med 2003;78(10 Suppl): S85-S87.
9. Cusimano MD, Rothman AI. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med 2003;78(10 Suppl):S88-S90.
10. Morrison LJ, Barrows HS. Developing consortia for clinical practice examinations: the Macy project. Teach Learn Med 1994;6(1):23-7.
11. Stern DT, Ben-David MF, Wojtczak A, Schwarz MR. Setting school-level international standards [abstract]. In: AMEE 2004 Conference; programme and abstracts. Edinburgh: AMEE; 2004:4.130.
12. Groot AD de. Vijven en zessen: cijfers en beslissingen: het selectieproces in ons onderwijs. 6e ongew. dr. Groningen: Tjeenk Willink; 1970.
13. Dousma T, Horsten A, Brants J. Tentamineren. 3e dr. Groningen: Wolters-Noordhoff; 1997.
14. Berkel H van, Bax A, editors. Toetsen in het hoger onderwijs. Houten [etc]: Bohn Stafleu Van Loghum; 2002.
15. Cohen-Schotanus J. Effecten van curriculumveranderingen: studiewaardering, studeergedrag, kennis, studiedoorstroom in een veranderd medisch curriculum [dissertation]. Groningen: [S.l: s.n]; 1994.

De auteurs:
Mw. dr. J. Cohen-Schotanus, hoofd Onderzoek en Innovatie Medisch Onderwijs, Faculteit der Medische Wetenschappen, Universitair Medisch Centrum Groningen.
Mw. dr. J. Schönrock-Adema, medewerker Onderzoek en Innovatie Medisch Onderwijs, Faculteit der Medische Wetenschappen, Universitair Medisch Centrum Groningen.
Prof. dr. A.J.J.A. Scherpbier, directeur Onderwijsinstituut, Faculteit der Medische Wetenschappen, Universiteit Maastricht.

Correspondentieadres:
Mw. dr. J. Cohen-Schotanus, Onderzoek en Innovatie Medisch Onderwijs, FMW/UMCG, A. Deusinglaan 1, 9713 AV Groningen, tel: 050-3632884, fax: 050-3633865, j.cohen-schotanus@med.umcg.nl.

Summary

Introduction: *Test results of groups of students vary within disciplines and between disciplines. This may be due to variations in test difficulty. Some methods for setting test standards take account of test difficulty. There is, however, no such thing as a 'perfect' standard setting method. The quality of standard setting procedures depends on several factors. One of these factors is the acceptability of the outcomes to different stakeholders. We examined for different assessment methods in medical education which fail percentages are considered acceptable by stakeholders.*
Method: *The participants of the Annual Conference of the Dutch Association for Medical Education (NVMO) were asked to indicate the fail percentages they considered acceptable for nine different assessment methods. Within the respondents we distinguished a subgroup of 'medical education experts'.*
Results: *The highest (26%) and the lowest (10%) fail percentages considered acceptable by the respondents were those for written tests in the preclinical phase and for the assessment of professional behaviour in the clinical phase, respectively. The group of 'medical education experts' were somewhat milder in what they considered acceptable.*
Conclusion: *The conference participants indicated what they considered to be acceptable standards for different assessment methods. The findings are in accordance with those of a comparable procedure among an international panel of medical education experts. Medical schools are advised to discuss the upper boundaries of acceptable fail percentages and to indicate what action is to be taken if these boundaries are exceeded. (Cohen-Schotanus J, Schönrock-Adema J, Scherpbier A.J.J.A. Which fail percentage is considered acceptable for different assessment methods? The opinion of participants of the Annual Conference of the Dutch Association for Medical Education (NVMO). Dutch Journal of Medical Education 2005;24(4):184-189.)*